(12) UK Patent Application (19) GB (11) 2 100 899 A

(21) Application No 8212846
(22) Date of filing 4 May 1982
(30) Priority data
(31) 8113829
(32) 6 May 1981
(33) United Kingdom (GB)
(43) Application published 6 Jan 1983
(51) INT CL³ G06F 3/023
(52) Domestic classification G4H 13D KU
(56) Documents cited GBA 2057973 GBA 2062916
(58) Field of search G4H B6F
(71) Applicants Li Jinkai, 330 Northeastern Building, Beijing Normal University, Beijing, China, Wei-Zang Chien, Beijing Qing Hua University, Beijing, China, An Qichun, 38 Sanlihe, Beijing, China, Wong Kam-Fu, 1211 Wu Sang House, Kowloon, Hong Kong
(72) Inventor Li Jinkai
(74) Agents Lloyd Wise, Tregear and Co., Norman House, 105—109 Strand, London WC2R 0AE

(54) Encoding ideographic characters

(57) Ideographic (e.g. Chinese) characters are encoded in numerical notation by allotting a different one of a number of digits to the possible different strokes or optionally combinations of strokes in the character and the order of the digits corresponds to the relative graphical positions of the strokes, highest first, then next highest and so on unless two or more strokes are on the same level in which case their order is from left to right. Optionally for more complicated characters the total number of digits in the code is limited by omitting strokes which are uncharacteristic or unnecessary to distinguish that character from any other. Such a method is relatively simple and can largely be understood even by someone with only a very limited knowledge of Chinese etc. writing. The method of the invention is useful, for example, for inputting Chinese etc. characters into a computer or teletypewriter and in preparing a dictionary of the characters.

FIG. 1.

TABLE 1

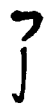
CODE	1	2	3	4	5	6	7	0
NAME	HORIZONTAL	VERTICAL	LEFT-SWINGING	POINT	LEFT- OR CLOCKWISE-TURN	RIGHT- OR ANTICLOCKWISE-TURN	POINT	SQUARE
SHAPE OF INDIVIDUAL STROKES	— /	|	) ,	`	⁊	L l	+ ×	□ □
CHARACTERISTIC	⟶	↓	↙	↘	↷	↻	+	□

GB 2 100 899 A

The drawings originally filed were informal and the print here reproduced is taken from a later filed formal copy.

FIG. 1.

TABLE 1

CODE	1	2	3	4	5	6	7	0
NAME	HORIZONTAL	VERTICAL	LEFT-SWINGING	POINT	LEFT- OR CLOCKWISE-TURN	RIGHT- OR ANTICLOCKWISE-TURN	POINT	SQUARE
SHAPE OF INDIVIDUAL STROKES								
CHARACTERISTIC								

2100899

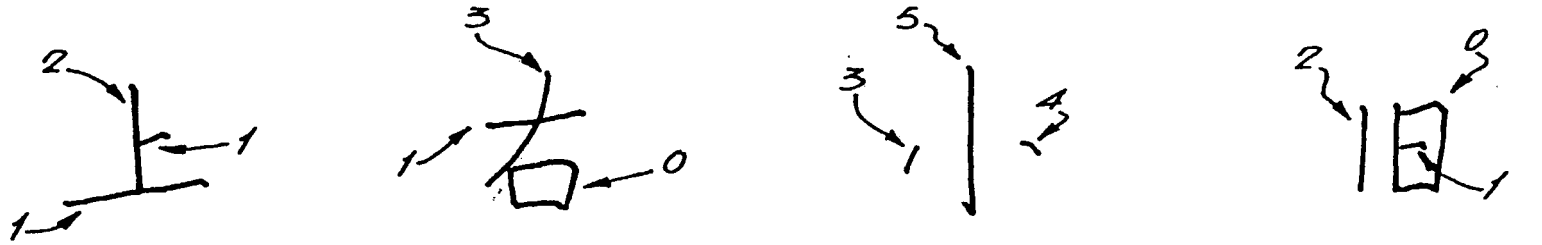
2100899

FIG. 1A.

TABLE 1A

CODE	1	2	3	4	5	6	7	0
DIRECTION OR SHAPE	→	↓	↙	↘	↻	↺	+	□
STROKE / SHAPE OF INDIVIDUAL STROKE / EXAMPLE OF STROKE	一 三	丨 中	丿 人	丶 主	乛 尸	乚 七	十 土	口 旦
	丿 刀	丨 五	丿 千	丶 入	亅 小	乚 氏	乂 凶	冂 目
		丨 真	丿 儿	丶 冗	乛 又	𠂉 允	乂 义	口 国
				丶 之	𠃌 凸	𠂉 凹	十 木	
				丶 心	乛 刃	乚 巴		
					𠃌 也	丶 心		
					𠃌 冗	丶 戈		
					丿 犯	𠂉 巡		
					𠃌 迁	乙 飞		
					𠃌 队	乙 九		
					𠃌 乃	乙 乙		
					𠂆 专	𠃌 认		
					𠂆 与			
					𠂆 鼎			

FIG. 2.



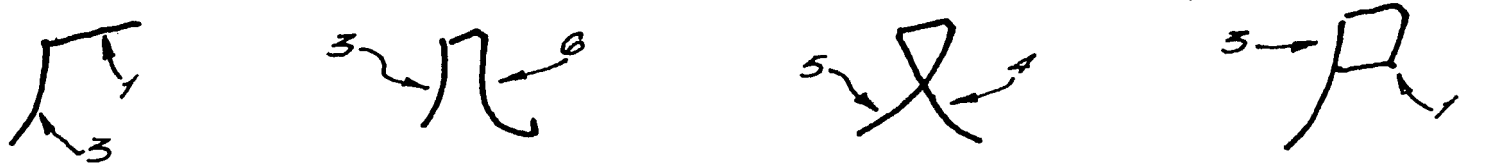
HIGHEST TO LOWEST
211

HIGHEST TO LOWEST
310

HIGHEST THEN LEFT TO RIGHT
534

LEFT TO RIGHT
201

FIG. 3.



13 (NOT 31)

36 (NOT 63)

54 (NOT 45)

531 (NOT 351)

FIG. 4A.

FIG. 4B.

FIG. 5.

2100899

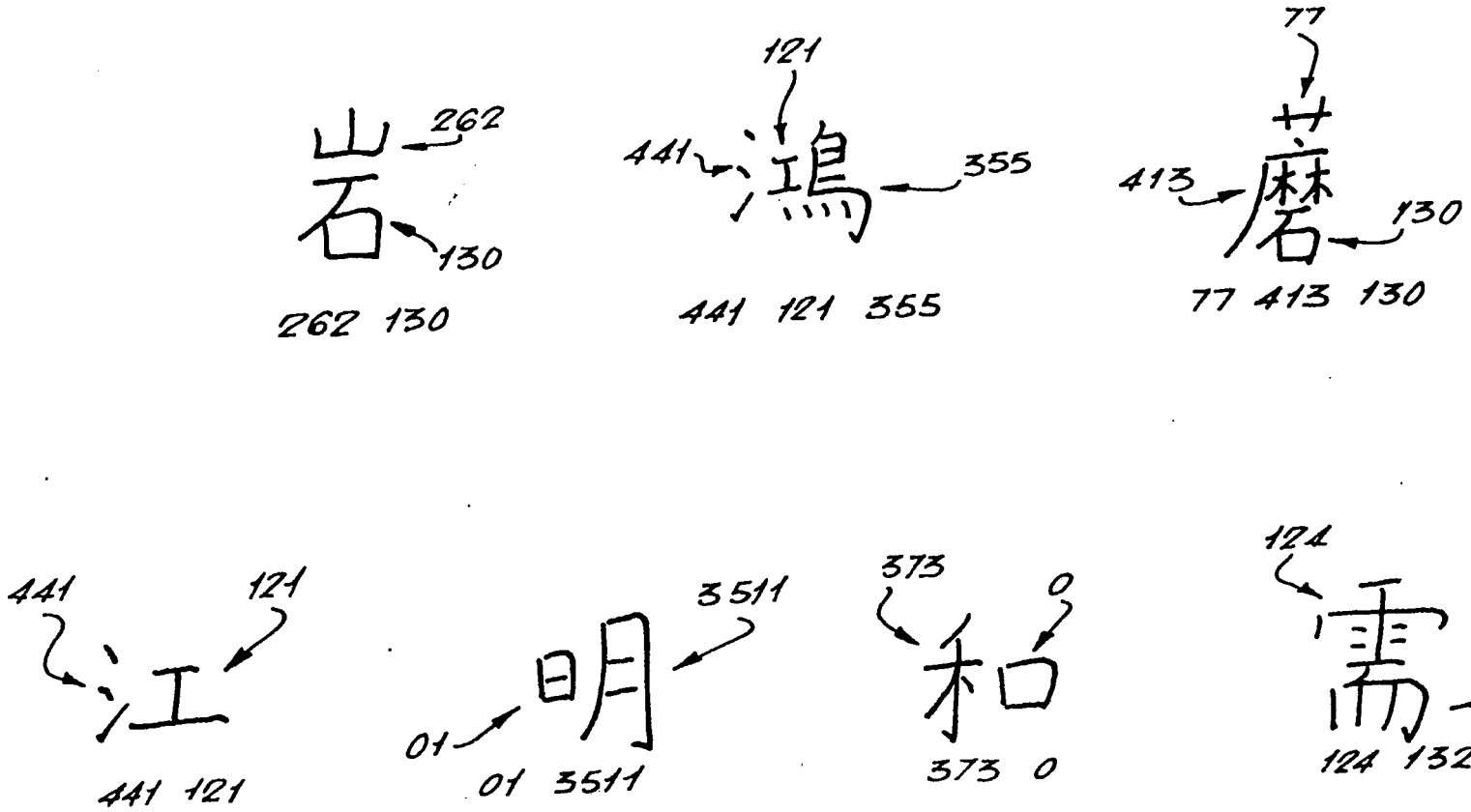
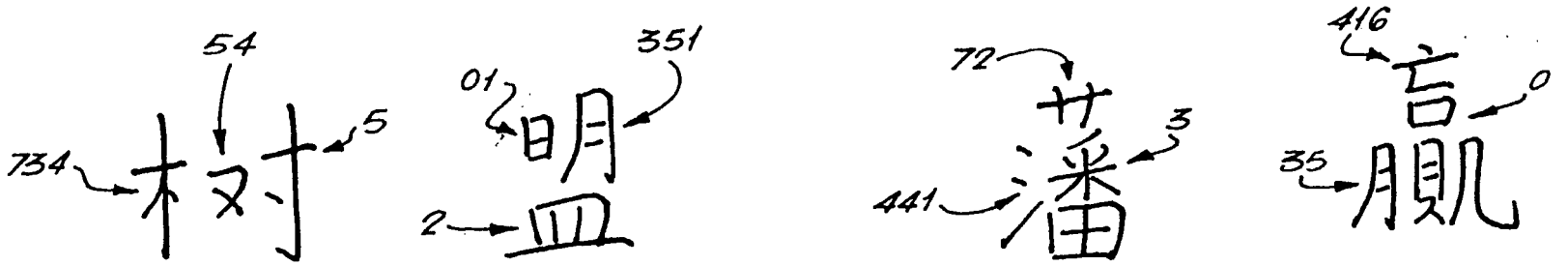


FIG. 6.

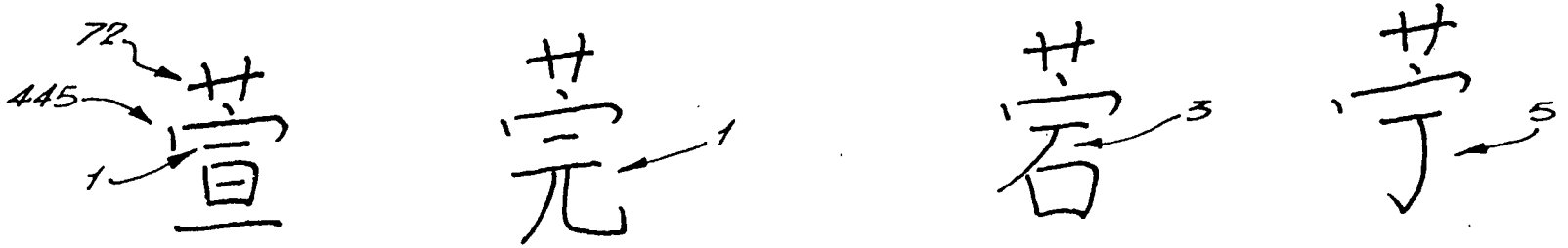
2100899

FIG. 7.



734 54 5 01 351 2 72 441 3 416 0 35

FIG. 8.



724451-0 724451-1 724451-3 724451-5

2100899

2100899

FIG. 9.

FIG. 10.

FIG. 11.

SPECIFICATION

System for encoding Chinese characters

This invention relates to a system for encoding Chinese characters so as to indentify the character in a form in which *inter alia* it can be fed into an electronic computer, telex or teleprinter machine or the like or it can be used to provide an index of Chinese characters.

Many attempts have been made in the past to develop a system for encoding Chinese characters but all such systems have proved complex and/or have needed the use of a large keyboard and/or involve a long code. In addition, such systems have tended to be complex and so any user of the system needs special and quite long training in order to be able to use the system.

It is therefore an object of the present invention to provide a simple encoding system for Chinese characters and in particular a system whereby the character can be encoded in a form suitable for simple and quick entry into an electronic computer.

According to the invention, there is provided a method of encoding Chinese characters in numerical notation in which the possible different strokes, and optionally the more frequently used combinations of strokes, of a Chinese character are allotted a different one of a number of digits, and a character is provided with a number code whose digits correspond to the strokes and optionally combinations of strokes in the character and whose order of digits corresponds to the relative graphical positions of the strokes and optionally combinations of strokes in the character so that, for example, the first digit corresponds to the highest, the second digit to the next highest and so on, unless two or more strokes and optionally combinations of strokes are on the same level in which case their order is taken as from left to right, and optionally in the case of the more complicated characters the total number of digits allocated is limited by omitting those strokes which are uncharacteristic or unnecessary to distinguish the overall character from any other.

Such a system has the advantage of being relatively simple and in addition the system can be grasped very quickly even by someone not conversant with Chinese characters and the method of writing Chinese characters. Thus each character is made up of a string of digits giving a unique number code to be fed into a computer or the like.

There are in the ultimate four different strokes which are used to make up all Chinese characters, namely horizontal, vertical, left-swinging and right-swinging. Therefore, each of these could be allotted a digit, e.g. 1, 2, 3 and 0 and this would be very convenient since the four digits would readily fit in two bits within the binary system basis of computers. With only four strokes, however, the number code for characters will be relatively long and this is often undesirable.

Therefore, certain combinations of strokes are preferably allotted their own digit. By choosing eight different categories of strokes, there are considerable advantages in simplifying the length of the digit code without making the code too complex. One could, however, also have, say, 16 categories since both 8 and 16 fit fully into three or four bits in the binary system so making the best use of a computer. Numbers other than binary numbers are possible, e.g. 5, but less efficient.

Therefore, according to a highly preferred embodiment of the invention, eight digits are chosen and in Table 1 forming Figure 1, the various strokes or combinations of strokes are listed together with one simple digit code which can be used. This code, for convenience, is shown as made up of the numbers 1 to 7 and 0 but of course the actual digits allotted to each stroke need not be the same as shown in the Table. More detailed illustrations of the strokes are allotted particular numbers and examples of these various strokes in use are given in Figure 1a.

By way of example, one can refer to the simple characters shown in Figure 2 beneath which are shown the resulting number codes according to the invention, the Roman numerals being used to show the order of the strokes.

There will be occasions when the point of initiation of two strokes is the same and so the order of the digits for a character ambiguous. In such cases this ambiguity can quickly and easily be resolved by selecting as the first digit the stroke whose termination point follows the graphic order highest first and then left before right. Examples of characters where this ambiguity would arise and the resulting number codes for those characters are shown in Figure 3.

The use of a graphic or positional rule according to the invention to decide on the order of the digits rather than a rule based on the order of writing the strokes, has the advantage of greatly reducing the chance of two characters having the same digit code. By way of example, reference can be made to the two Chinese characters shown in Figures 4A and 4B. Both of these characters could be allotted the digit code 1134 if the digit code is chosen as in some prior methods according to the normal writing order of the strokes. This ambiguity, however, can readily be avoided by choosing the order according to the invention and so by taking the highest stroke or combination of strokes first. In this example, the character shown in Figure 4B remains 3114 whilst the character shown in Figure 4A becomes 1314. Also the graphic rule used according to the invention has the advantage that it can be understood by someone who is not conversant with Chinese characters or the order in which the strokes are written.

A further advantage of the invention is that a graphic rule adopted according to the invention to decide on the order of digits is unique whereas a rule based on the order or writing the strokes may not be since there is no really universal rule for the order of drawing out a Chinese character.

By following the system according to the

invention, there will be very few characters which cannot be given a single number code. By way of example, one might refer to the Chinese character shown in Figure 5. This could be encoded as any one of the groups of digits 373, 733 or 337. Since this will only arise very rarely, one can accept the situation and, for example, index this character under all possibilities or in, say, the use of a computer, suitable software can be programmed in to accept any of the possibilities as representing the required character.

For simplicity hereinafter, references to "strokes" are to be construed as meaning either individual stroke or combination of strokes unless the context specifically requires otherwise.

As can be seen in the tables, for some digits there are two or more possibile strokes. This does not lead to confusion in the system according to the invention. Thus, when a Chinese character is made up of a number of digits according to the invention, it will be possible for a person conversant with the Chinese characters to know which of the two possible strokes is the only one intended from the remaining digits of the code for that character. For example, in the case of digit 7, of the two possible orientations of the cross only one will be appropriate to complete a character when the remaining digits allotted to the character have been converted to their appropriate strokes. Of course, if, say, 16 digits are chosen instead of 8, this use of two possible strokes for some of the digits can be reduced but as explained above, this will not usually be necessary and an increase in the number of possible digits over 8 will tend significantly to increase the length of the number code.

An advantage of the invention is that only four, or in a more preferred embodiment eight, keys in a keyboard are required for the entry of any Chinese character into, say, a computer. This means that the keyboard can be relatively small and compact. In addition, if a separate keyboard is not provided, there is also the advantage that digit keys will already exist on conventional computer, calculator and typewriter keyboards and so these existing keys can be used together with a function key.

Because only a very limited number of keys and digits are necessary and because the rule defining the order of the strokes is based on a simple visual test, the input speed for characters into, say, a computer can be quite high and in fact it is believed that after a few hours' practice, an operator with average typing skill could input characters at the rate of as many as 4,600 characters per hour.

The method of the invention is not only useful in the input of Chinese characters into computers and the like but can be used to index a Chinese dictionary. Thus, despite the existence of the Chinese language for thousands of years, there still remains a problem as yet not satisfactorily solved of indexing Chinese dictionaries. This could be solved according to the invention and the dictionary or indexes to books, listed according to the digits allotted as described above. Thus, all those characters starting with digit 1 could be grouped together and then sub-divided successively according to the second, third, fourth and so on digits allotted. In this respect, the order in the index could be analogous to the order of classification of the subject matter of books and the like according to the Dewey Decimal Classification System.

The encoding system according to the invention is simple. There is no necessity to inspect leading parts count the number of strokes, keep the writing order in mind or analyse four-corners. It should prove very easy to become familiar with the system and to use it even for a user whose mother tongue is neither Chinese or Japanese. Furthermore, if the system is adopted to index dictionaries, then anyone who can use a dictionary can be easily trained to operate an input keyboard processing Chinese character information.

Although reference has been made above to Chinese characters, the invention is equally applicable to the encoding of ideographic characters other than Chinese characters, e.g. Japanese and Korean characters. Thus, references herein to "Chinese characters" are to be construed broadly to encompass ideographic characters of languages having a similar common source.

According to a preferred embodiment of the invention where there are complex Chinese characters, these can be allotted a number code according to the invention by dividing the complex character into its known roots. Thus, anyone conversant with Chinese characters would readily know the two or three or more different roots from which a more complex character is built up. Then the character is allotted a digit code according to the invention by giving a code to the first root, a code to the second root and so on. If the character is exceptionally complex and contains more than three roots, it is usually sufficient to give number codes to no more than three roots, preferably the first, second and third, since that will usually be sufficient to identify the character uniquely.

The order of the number codes for each root in the complex character is preferably selected so as to be the same as the graphic order chosen for the strokes in a character or root. Thus, the number code for the root which is highest is placed first, followed by the next highest and so on and in the event that two or more roots are at the same level or approximate level then the root to the left is placed before a root to the right.

By way of example, reference is made to Figure 6 which shows a number of more complex characters and the number codes allotted according to their various roots.

According to a further preferred feature of the invention, each root in a more complex Chinese character is given a limited number of digits, preferably 3. This becomes possible because a complex character can usually be identified uniquely from a maximum of three strokes in each of the roots. Thus, whilst to give an exhaustive number code to each root might require, say, five

digits anyone conversant with the language would readily appreciate that two or perahps three digits for each root will be enough to identify the unique roots and combinations of roots making up the complex character. Thus, one can omit from the characterisation of each root those strokes which are characteristic and superfluous. In the list set out in Figure 6, this shortening has, for example, been made in connection with the root given the digit 355 above but was not necessary or desirable with the two root character 01 3511 since the total digit code was not particularly long.

Most preferably where a complex character has three or more roots, each root is limited to a digit code of no more than three digits and the whole character, where made up of a number of roots or not is limited to a total of six digits so that a minimum number of entries are made via a keyboard. This will normally be possible since such complex characters can usually be identified readily since there are a relatively limited number of such complex characters. Thus in the examples of characters shown in Figure 7, the overall digit code has been limited to six digits and each character shown can be uniquely identified by that six digit code.

There will be some rare occasions where a number of characters will have the same six number digit code and examples of these situations are shown in Figures 8 and 9. In order to distinguish the characters in such situations it will be necessary to use a seven digit code, and very exceptionally an eight digit code. Preferably in these circumstances the seventh or seventh and eighth digits chosen are the next following distinguishing stroke or strokes of the character and this is illustrated in Figures 8 and 9.

By way of example, there are about 6,700 characters in common use in the Chinse language and with a six-digit code, only about 1.3% of these cannot be uniquely identified. If a seventh digit is used, as described above, then only 0.6% can then not be uniquely identified.

Shortening of the digit code can be applied to the whole character itself if the character is made up of only a single root of which only a limited number of strokes are characteristic. To take, for example, the Chinese character shown in Figure 10, this would be encoded according to the invention with the number code 212543 but as there is no other word whose number code begins with 212, for this character only the three-digit code 212 would be enough. To simplify the rules which the operator has to learn, one could, of course, build into a computer appropriate software which would identify that Chinese character by the number code 212 so that the computer simply ignores the remaining characters 543 in its processing and in its print-out.

One can additionally simplify the number codes prepared according to the invention for a selected number of Chinese characters and words which appear with great frequency and such words can be given a unique one or two or three number code. For example, the character shown in Figure 11 is a word very frequently used in Chinese and only one stroke, say, the left swinging stroke at its upper-left corner may be taken to represent the whole character making the code of the world a mere "3".

As noted above, the system of the invention can be used with existing keyboards and in addition could be used to transmit Chinese characters via existing telex machines or teleprinter machines provided the operators at both ends of the telephone wire know the code between the character and the digits and a code book listing the characters exhaustively is not required so that the time for consulting such a code book can be avoided.

CLAIMS

1. A method of encoding Chinese characters (as herein defined) in numerical notation in which the possible different strokes, and optionally the more frequently used combinations of strokes, in a Chinese character are allotted a different one of a number of digits, and a character is provided with a number code whose digits correspond to the strokes and optionally combinations of strokes in the character and whose order of digits corresponds to the relative graphical positions of the strokes and optionally combinations of strokes in the character, and optionally in the case of more complex characters the total number of digits allocated is limited by omitting those strokes or which are uncharacteristic or unnecessary to distinguish the overall character from any other.

2. A method as claimed in Claim 1 in which the strokes which are allotted digits are horizontal, vertical, left-swinging and right-swinging and the corresponding digits allotted are 1, 2, 3 and 0.

3. A method as claimed in Claim 1 in which the strokes which are allotted digits and the corresponding digits are as shown in Table 1 of Figure 1.

4. A method as claimed in any preceding claim in which the order of the digits in the number code of a character is such that the first digit corresponds to the highest stroke or optionally combinations of strokes, the next to the next highest and if two or more strokes or optionally combinations of strokes are on the same level their order is taken as from left to right.

5. A method as claimed in any preceding claim in which for complex characters formed of a number of roots, a number code is determined for each root and then the order of the number codes for the root to give the number code for the whole character is determined by the same relative graphical positions as the strokes in the roots.

6. A method as claimed in any preceding claim in which the number code for a character is limited to no more than nine digits.

7. A method as claimed in Claim 6 in which the number code for a character is limited to no more than six digits.

8. A method of encoding Chinese characters in numerical notation substantially as herein described.

9. A dictionary of Chinese characters in which the characters have been encoded by a method as claimed in any preceding claim and the order of the codes set out in the dictionary is such that all those starting with the same digit have been grouped together and then sub-divided successively according to the second, third, fourth and so on digits.

10. A computer which has been programmed to print out Chinese characters according to the method of encoding as claimed in any of claims 1 to 8.

Printed for Her Majesty's Stationery Office by the Courier Press, Leamington Spa, 1983. Published by the Patent Office 25 Southampton Buildings, London, WC2A 1AY, from which copies may be obtained.